R E P O R T R E S U M E S

ED 018 153 56 FL 000 147

MEASUREMENT OF SPEAKING SKILLS IN ELEMENTARY LEVEL SPANISH INSTRUCTION. DENVER-STANFORD PROJECT ON THE CONTEXT OF INSTRUCTIONAL TELEVISION, REPORT NO. 9.

BY- ANDRADE, MANUEL AND OTHERS
STANFORD UNIV., CALIF.INST. FOR COMMUN. RES.

REPORT NUMBER NDEA-7A-354 PUB DATE JUL 63
DENVER BOARD OF EDUCATION, COLO., SCH. DIST. NO. 1
GRANT OEG-7-14-1380-083
EDRS PRICE MF-$0.25 HC-$1.24 29P.

DESCRIPTORS- *SPANISH, *FLES, *SPEECH SKILLS, *TESTING, *TEST CONSTRUCTION, STATISTICAL ANALYSIS, TESTING PROBLEMS, TEST RELIABILITY, STUDENT TESTING, TEST VALIDITY, TEST INTERPRETATION, INSTRUCTIONAL TELEVISION, TEACHING METHODS, AUDIOLINGUAL METHODS, GRADE 5, GRADE 6, LANGUAGE INSTRUCTION, ANALYSIS OF VARIANCE, DENVER STANFORD PROJECT, DENVER, STANFORD,

SINCE NO TESTS OF SPANISH SPEAKING ABILITY AT THE ELEMENTARY LEVEL WERE AVAILABLE WHEN THE PROJECT BEGAN (1960), IT DEVISED THREE CAREFULLY CONSTRUCTED ITEMS TESTING THE SEPARATE ASPECTS OF THE SPEAKING SKILL--PHONETIC ACCURACY, STRUCTURE, AND EASE AND NATURALNESS OF EXPRESSION. A RANDOM SELECTION OF FIFTH-GRADE PUPILS WERE TESTED INDIVIDUALLY, AND THEIR RESPONSES RECORDED ON MAGNETIC TAPE AND EVALUATED SEPARATELY BY AT LEAST TWO PERSONS. TEST ITEMS WERE DESIGNED TO BE EXPLICIT, AND TO REFLECT THE COURSE CONTENT AS CLOSELY AS POSSIBLE. SINCE NO OUTSIDE CRITERIA WERE AVAILABLE, THE BOOKLET OFFERS A DETAILED MATHEMATICAL PRESENTATION OF THE STATISTICAL TECHNIQUES APPLIED TO THE COMPLEX PROBLEMS OF VALIDITY AND RELIABILITY. A MAJOR REVISION, BASED ON STATISTICS COMPILED FROM THE FIRST TEST, AND FURTHER DEVELOPMENTS OF THE TEST ARE DESCRIBED IN DETAIL. THE TESTS HAVE BEEN FOUND SATISFACTORY BOTH AS CRITERIA OF PUPIL ABILITY AND IN DIFFERENTIATING BETWEEN TEACHING METHODS. FOR COMPANION DOCUMENTS SEE FL 000 813, FL 000 820, AND FL 000 821. (RW)

ED018153

MEASUREMENT OF SPEAKING SKILLS IN ELEMENTARY LEVEL SPANISH INSTRUCTION

• MANUEL ANDRADE
• JOHN L. HAYMAN, JR.
• JAMES T. JOHNSON, JR.

FL 000 147

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Denver Public Schools • Stanford University

RESEARCH ON THE CONTEXT OF INSTRUCTIONAL TELEVISION

The research reported herein was supported by a grant from the United States Office of Education, Department of Health, Education, and Welfare.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

MEASUREMENT OF SPEAKING SKILLS IN ELEMENTARY LEVEL SPANISH INSTRUCTION

by

Manuel Andrade
John L. Hayman, Jr.
James T. Johnson, Jr.

DENVER-STANFORD PROJECT
ON THE CONTEXT OF INSTRUCTIONAL TELEVISION

****

School District Number One
City and County of Denver
Denver, Colorado

Institute for Communication Research
Stanford University
Stanford, California

Report Number 9
July 1963

FL 000 147

SUMMARY

The Problem

The Denver-Stanford project is involved with teaching Spanish to fifth and sixth grade pupils in the Denver Public Schools, and one of its concerns has been the evaluation of these pupils' abilities to speak Spanish. When the project began in 1960, no tests of Spanish speaking ability at the elementary school level were available. Project personnel therefore began development of speaking tests.

Results

Careful review of relevant literature led to the conclusion that the speaking skill could be broken down into three distinct aspects: the ability to pronounce Spanish sounds properly; the ability to compose Spanish sentences orally, using correct syntax and grammar; and the ability to communicate in Spanish with ease and naturalness. To measure these separate aspects, speaking tests, composed of phonetic accuracy, structure, and fluency sections, were constructed.

The tests were administered by project personnel to pupils selected randomly. Each pupil's performance was recorded on magnetic tape, and each was in turn evaluated independently by at least two persons.

Both composite and rater reliabilities were computed in statistical evaluation of the tests. Since each test part measured a separate aspect of the speaking skill, internal validity varied inversely with composite reliability. Consequently, a low alpha coefficient, the measure of composite reliability, was sought. Rater reliability, on the other hand, reflected the extent to which similar scores were assigned each pupil by the separate evaluators, and, therefore, a high figure was sought.

The development process revealed several points to be considered in constructing a foreign language speaking test. If the test parts are really to reflect different aspects of the speaking skill, they must be evaluated separately, and the evaluator must be careful not to be influenced by performance on one section when scoring another. A two- or three- point scoring scale, with each scale position defined by a specific behavioral element, seems desirable. Finally, each test part should produce about the same mean score and about the same variance to weigh equally in the total test score.

The development was completed during the 1960-61 school year, and the tests have been used in subsequent years and found to be satisfactory.

MEASUREMENT OF SPEAKING SKILLS IN ELEMENTARY LEVEL SPANISH INSTRUCTION

by

Manuel Andrade
John L. Hayman, Jr.
James T. Johnson, Jr.

The Denver Public Schools and Stanford University's Institute for Communication Research are currently engaged in a joint research project on the context of instructional television. The purpose of the project is to learn how instructional television can best fit into the total teaching situation. A substantial amount of research has established that television is a very effective teaching medium. Ways of combining it with other educational activities must now be considered, and the Denver-Stanford project is a beginning effort in this direction. Kenneth E. Oberholtzer is principal investigator for the Denver Public Schools and Wilbur Schramm is principal investigator for Stanford University. This is one of a number of project progress reports.

The Problem

The primary purpose of the Denver-Stanford project is to explore the context of instructional television and to improve the effectiveness of instruction by changes in context. Elementary school Spanish was chosen as the subject matter to be used throughout the project. Therefore, though the teaching of Spanish per se is secondary to project aims, it is essential to the welfare of pupil participants that the best teaching methodology in this field be utilized.

In line with the latest findings on language instruction and the recommendations of those associated with the Foreign Language in the Elementary Schools' (FLES) program, the audio-lingual approach has been used exclusively during the first year of instruction (fifth grade), and it plays a major

role during the second year (sixth grade) although reading and writing are introduced then.

The first skill which pupils must acquire in this approach is listening comprehension -- the ability to understand what is said in the second language. The second skill is the ability to speak in the second language and to carry on meaningful communication. A facility in both listening and speaking must be acquired before the child begins to read and write (Brooks, 1960, pp. 119-132).

Measurement is necessary, of course, both to evaluate experimental procedures and to determine if the general aims of language instruction are being satisfied. Five listening comprehension tests for administration via television have been developed by project personnel, and this development -- which was relatively straight forward -- is described in a previous progress report (Andrade, Hayman, and Johnson, 1961).

Considerable effort has also gone into the development of speaking tests. This type of measurement is much more complex, however, for, as has been observed elsewhere, "speaking ability presents the most difficult problem in (foreign language) testing" (State Department, 1958, p. 14). Huebener suggests the basis of some of these problems as follows:

> Speaking ability is the most difficult phase of a foreign language to teach and to acquire.
>
> This ability is least likely to be retained, for it depends on constant practice.
>
> It is difficult to teach because it requires unusual resourcefulness, skill, and energy on the part of the teacher. Teaching ability cannot be acquired through a textbook (Huebener, 1959, p. 8).

Huebener makes it clear that considerable experience and training is necessary before a teacher can adequately teach pupils to speak in a second language. And certainly a teacher must be well qualified before he can validly judge the speaking performance of others. Even if the teacher is sufficiently skilled to evaluate speaking performances, however, test

administration is difficult. Keesee points out that, ". . . this skill (speaking ability) is measured only through providing an opportunity for the pupil to speak" (Keesee, 1960, p. 60). Handling pupils individually is at best a time consuming, exhausting process, and it requires painstaking care to assure similar test conditions for every subject.

The Denver-Stanford project currently has over 13,000 fifth and sixth grade pupils participating, with more than 350 teachers handling classroom activities. As in other localities, only a small per cent of teachers have the training and experience to qualify them as experts in Spanish. This means, therefore, that only a few of those in the project could validly and reliably handle the measurement of speaking skills. In light of the necessity for such measurement, this situation -- combined with the difficulties inherent in assessing the ability to speak -- has presented a real challenge to project personnel. This report describes the attempt to meet this challenge in the development of oral measuring instruments, and it discusses the use of these instruments in the project.

Development Criteria

Language Skills. According to MacRae, an audio-lingual language program at the elementary school level is built on the following learning experiences: "Hearing the new language in meaningful patterns, imitating the new sounds by rote, speaking the new language in meaningful situations, and recombining vocabulary thus acquired in class-originated oral experiences" (MacRae, 1957, p. 24). And, as MacRae says further, "The skills that boys and girls in the elementary grades may be expected to develop are closely related to the learning experiences we have just noted. . ." (Ibid., p. 25).

For testing purposes, these skills must, of course, be defined in terms of specific behavioral elements, and, again according to MacRae, they can

be defined as the pupil's ability:

> To speak Spanish with ease and naturalness and an acceptable unanglicized accent.
>
> To have developed the ability to listen carefully enough to retain and repeat new sounds.
>
> To become aware of the mechanics of speaking.
>
> To realize something of language structure, not grammar as such, but that words have different functions to perform as they are fitted together to express meaning.
>
> To have acquired by ear one of the most important characteristics of Spanish structure, the agreement of nouns and adjectives (Ibid.).

It seemed to project personnel that these abilities could be measured by a test composed of three distinct sections: phonetic accuracy, structure, and fluency. The phonetic accuracy section would test the pupil's ability to pronounce Spanish words properly and to repeat sounds, the structure section would test his ability to use correct syntax and grammar in orally composing Spanish sentences, and the fluency section would test his ability to communicate in Spanish with naturalness and ease.

Conditions of Administration. Administration of a test of language speaking skills presents special problems per se. Each subject must be allowed an opportunity to perform, but this performance cannot be in the classroom since administering the test in the classroom would favor those pupils who heard the items several times before their turn to be tested.

Furthermore, in at least the development and early validating stages of a subjectively scored measuring instrument, each pupil's performance should be evaluated by two or more persons working independently. The reasons for this will be apparent in the next section. The point here is that, though at least two evaluations are needed, having two or more evaluators present at the test administration would be undesirable because it would involve inefficient use of a considerable amount of project personnel time and because it would make independent evaluation difficult.

The solution to these problems seemed to lie in testing pupils individually

in a room with only the tester and pupil present. Recording the performance on magnetic tape would allow independent evaluation at a later date. And, finally, the individual testing situation would allow the tester greater control and would assure, within reasonable limits, similar testing conditions for all subjects.

Validity and Reliability. Validity was a difficult problem because of the lack of an outside criterion against which to compare obtained results. In the first place, project personnel were unable to locate a speaking test designed for elementary school Spanish. In the second, even if one were available, its adequacy, in terms of specific needs of the Denver Public Schools' Spanish program, would be questionable.

The test should be comprehensive, that is, it should be a representative sample of the course content. And, as Keesee has noted, "the pupils (should be) tested as nearly as possible in the manner in which they have been taught. . . . No complicated unfamiliar visual materials should be introduced in a test" (Keesee, 1960, p. 61).

The only alternative in this situation is to use construct validity, in which the test objectives are ". . . made so explicit that one can determine (without empirical demonstration) whether each answer to a test item is a behavior belonging in the class (of behaviors) in question" (English and English, 1958, p. 575). The test items were chosen jointly by several persons who were thoroughly familiar with course content and objectives. In making choices, this group kept in mind the need for comprehensiveness of the test as a whole and for preciseness in definition of individual behaviors sought.

The need for content validity and generally understood principles of testing necessarily restrict test content to course content. A test should be a representative sample of course content, and as such it will be comprehensive. A test must not go beyond course content, however. The behavioral

elements chosen for evaluation, then, were elements which had been taught.

Reliability had to be approached in a manner different from that normally employed. The split-half method could not be used because it requires that a test, by some means, be divided into parallel parts, and, according to Guilford, "To be parallel parts, . . . the subtests that compose the parts should have items of equal average difficulty, equal spread of difficulty, and equal item intercorrelation, and the same amount of time should be devoted to each" (Guilford, 1954, p. 377). These conditions would obviously be most difficult to satisfy in the proposed speaking test.

One appropriate method of estimating reliability under these conditions is to use Cronbach's generalized equation, which produces what Cronbach has named the "coefficient alpha." The formula for coefficient alpha is:

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma V_i}{V_t}\right)$$

where V_i = variance of part I of a test, the size not specified

V_t = variance of total scores

n = number of parts.

The alpha coefficient in this case will give a composite reliability, which reflects, among other things, the dispersions of the separate components of the test and the component intercorrelations. As Guilford states, "High intercorrelations of components detract from validity of the composite. Where validity is at stake for a composite, we would therefore not strive toward high composite reliability but the reverse" (Ibid., p. 393).

Another problem in reliability existed because of the subjectivity in evaluating results. In this case, the scorer as well as the test content contributes errors of measurement. According to Guilford, the preferred method of estimating rater reliability is to correlate scores assigned by different persons working individually (Ibid., p. 395). One approach to this problem is

offered through intraclass correlation, for which Ebel has given the following formula:

$$\bar{r}_{11} = \frac{V_p - V_e}{V_p + (k - 1)V_e}$$

where $\bar{r}_{11}$ = the mean reliability of ratings for one rater

V_p = variance for persons

V_e = variance for error

k = number of raters.

The reliability of the mean of k ratings for each person would be:

$$r_{kk} = \frac{V_p - V_e}{V_p}$$

The computation formulae for computing the appropriate variances are given on pages 396 and 397 of Guilford's Psychometric Methods (Ibid.).

Another approach to the problem would be to compute Pearson product moment correlation coefficients (r_{ab}), though this method would have the disadvantage of producing a separate figure for each pair of raters.

With all of the rater reliability estimates, the object is to produce coefficients as high as possible, that is, to produce maximum agreement among raters.

Reliability estimation for the type of test being developed, then, was approached in two distinct ways, each involving a different objective. For composite reliability, the objective was to minimize the reliability rating. For rater reliability, on the other hand the objective was to maximize the rating.

The First Trial Test

Make-up of the Test. A speaking test, with sections indicated in the previous section and based on trials with a few children, was constructed in

the fall semester of the 1960-61 school year. The test was administered at the end of the semester to a random sample of 130 fifth grade children who were taking first year Spanish and therefore in the research project.

In the phonetic accuracy part of the test, the tester spoke the Spanish sentence, "El hijo pequeño tiene un libro amarillo" (The small boy has a yellow book), and the subject then repeated the sentence. This sentence was designed to allow the pronunciation of all of the Spanish vowels and the consonants "ñ" and "ll" to be evaluated.

In the structure section, the tester asked the following five questions, and the subject was asked to respond in complete Spanish sentences.

1. ¿Cómo se llama usted? (What is your name?)
2. ¿Qué artículos de ropa usa un niño? (What articles of clothing does a boy wear?)
3. Dígame usted las partes del cuerpo. (Tell me the parts of the body.)
4. ¿Qué se pone usted en los pies? (What do you put on your feet?)
5. ¿Cuántos ojos tiene usted? (How many eyes do you have?)

This section was designed to allow evaluation of pronunciation, syntax, structure, extent of vocabulary, spontaneity of response, and appropriateness of response.

In the fluency section, a visual, which clearly showed members of a family, parts of the body, and articles of clothing, was displayed. The subject was instructed to tell all that he could about the picture in complete Spanish sentences.

Scoring the Test. The test was scored independently by two members of the project staff. The scorers went over several sample performances together so that their evaluation criteria would be as similar as possible. Then each went through the total group of performances without knowing what scores the other had assigned.

The scoring itself was accomplished with a standard rating sheet (Appendix A),

on which the pupil's performance on each test item was rated from excellent to poor on a five-point scale. The scorer would first listen to the complete performance and then listen to and evaluate the separate sections. If the scorer was uncertain as to the exact scale position to be marked for a particular item, he would consider the child's performance on the skill measured by this item in other parts of the test.

Test Reliability. The composite reliability for the test, as measured by the alpha coefficient, was:

$$\alpha = .740.$$

In light of the desire for validity and therefore low composite reliability, this alpha coefficient seemed too high. It would indicate, among other things, high intercorrelations between test parts. These intercorrelations plus the correlations of each test part and the total test with the first semester listening comprehension test (the measure of the understanding skill) are shown in table 1. Pearson product moment correlations are given in this table.

Table 1

CORRELATIONS BETWEEN SPEAKING TEST PARTS, TOTAL SPEAKING TEST, AND LISTENING COMPREHENSION TEST --FIRST SEMESTER 1960-61

Test part	Structure	Fluency	Total Test	Listening Comprehension Test
Phonetic Accuracy	.502	.487	.776	.445
Structure		.368	.908	.662
Fluency			.812	.740
Total Test				.714

Table 1 shows rather high intercorrelations between test parts, and it indicates that the structure section was doing about the same thing as the test as a whole. The correlations of test parts with the total are spurious, of course, because each part contributes to the total. To determine the

correlation between each part and the total, with the influence of that part removed, part-whole correlations were computed (McNemar, p. 164). The part-whole correlation of phonetic accuracy with the total was .534, of structure with the total was .653, and of fluency with the total was .687. Again the figures indicate high relationships, probably higher than would be expected if the test parts were really measuring different aspects of the speaking skill as desired. In this respect, a need for improvement was definitely indicated.

Rater reliabilities for the test were surprisingly high. Table 2 shows rater reliabilities for each part of the test and the total test in terms of the three coefficients discussed earlier. As mentioned previously, two raters were used. Rater reliability, then, seemed to be satisfactory.

In table 2, $\bar{r}_{11}$ and r_{ab} are the same in each comparison. This suggests that the product moment correlation, r_{ab}, is a special case of $\bar{r}_{11}$ where only two raters are involved. As proved in Appendix C, this is only true if the variance of scores assigned by both raters is the same. Test part means and variances for each rater are shown in table 3. Though the means differ somewhat, variances are indeed quite similar, and this explains the similarity between $\bar{r}_{11}$ and r_{ab}.

Table 2

RATER RELIABILITY ON THE
1960-61 FIRST SEMESTER SPEAKING TEST

Test Part	Reliability Coefficient r_{kk}	$\bar{r}_{11}$	r_{ab}
Phonetic Accuracy	.966	.935	.935
Structure	.971	.943	.943
Fluency	.861	.756	.756
Total Test	.976	.952	.952

Table 3

MEANS AND VARIANCES FOR EACH RATER
ON THE 1960-61 FIRST SEMESTER SPEAKING TEST

Test Part	Rater	Mean	Variance
Phonetic Accuracy	A	16.563	33.989
	B	17.126	34.152
Structure	A	18.650	65.675
	B	19.116	66.325
Fluency	A	12.971	22.801
	B	8.283	22.222
Total Test	A	48.582	234.182
	B	44.582	238.424

Though the reliabilities were quite satisfactory overall, table 3 shows that rater B gave somewhat higher scores on the average for phonetic accuracy and structure than rater A, while rater A gave higher scores on the fluency section. The raters could not be expected to give identical scores in each section, of course, nor would the differences between them always be in the same direction. Under ideal conditions, however, differences would be at the chance level, which is definitely not the case in the fluency section.

The Second Trial Test

Make-up of the Test. The test make-up was revised during the second semester in light of statistics compiled on the first trial test and of experiences of the evaluators in both the administration and scoring of the first test. In addition, the new test covered course content from the complete year rather than just the first semester.

The evaluators found in administering the phonetic accuracy section of the first test that the sentence to be repeated was too long, and that many pupils consequently forgot a word or two. This caused them to lose points even if

the sounds they remembered and did pronounce were quite accurate. Therefore, a shorter sentence, "Es una señorita" (It is a young lady), was used. As before, the tester spoke the sentence, and the subject repeated it. This sentence allowed evaluation of all of the Spanish vowels.

In the structure section, the evaluators found in the first test that the rather general questions asked allowed too many possible valid responses. This section therefore was more rigidly structured in order to predetermine and limit the possible responses. The tester supplied the subject with vocabulary, not in syntactically correct order, needed to construct a sentence. Each word was established independently with visuals. A picture of a man was used to establish the noun and article, "el padre" (the father). A picture of a boy wearing shoes was used to establish the verb and object, "usa zapatos" (wears shoes). And two strips of black paper were used to establish the plural of the adjective, "negros" (black). Then the subject was shown a picture of a man wearing black shoes and asked to construct a sentence from the established vocabulary which would describe the visual. The correct response would be, "El padre usa zapatos negros" (The father is wearing black shoes). The vocabulary used has been thoroughly covered in the course. It was specifically established here so that only the child's ability to arrange the word in correct order would be measured.

As in the previous section, the evaluators felt that the stimuli provided in the fluency section did not structure the possible responses sufficiently. Therefore, instead of four visuals which the subject was asked to describe in complete Spanish sentences, four different tasks were required. One of these involved a visual to be described, as before. The others involved: (1) asking the question, "¿Cómo se llama usted?" (What is your name?), for which there is only one answer; (2) handing the child an apple and asking, "¿Qué tiene usted en la mano?" (What do you have in your hand?); and (3) displaying a visual and

asking, "¿Quiénes pagan for los comestibles?" (Who pays for the groceries?).

The test was administered, on an individual basis, to 200 randomly-selected pupils at the end of the semester.

Scoring the Test. To preserve the integrity of each test part, that is, to make each test part reflect a specific aspect of the speaking skill and not be influenced by other test parts, the parts were scored separately. The evaluator would listen to the phonetic accuracy section, for example, as many times as he liked in making his judgment, but he would not listen to a succeeding section until scores for the one in question were assigned. And he would try not to be influenced in his present evaluation by the child's performance in preceding sections, though preserving such independence of thought in actual practice is difficult.

In addition, the evaluators felt that the five-point scale used in the first test demanded finer discrimination in judgment than could validly be made. Consequently, the five-point scale was abandoned and three- and four-point scales were adopted. In another refinement, each scale position was precisely defined, as opposed to the first test in which each scale represented a range from very poor to very good without any specific behavioral element indicating a certain scale position.

In the phonetic accuracy section, each vowel was rated as follows:

2 = accurate reproduction,

1 = inaccurate reproduction.

0 = no production.

In the structure section, the scale scores were:

2 = complete sentence, syntactically correct

1 = incomplete sentence, syntactically correct

0 = sentence not syntactically correct or no measurable response.

In the fluency section, the scoring was as follows:

. In answering the question, "¿Cómo se llama usted?"

3 = "Me llamo _____" or "Yo me llamo _____"

2 = "Se llamo _____" or name only

1 = "Se llamo es _____," "Me llamo es _____," or any other combination in which the name is stated

0 = inappropriate response or no response.

. In answering the question, "¿Qué tiene usted en la mano?"

3 = correct response, given naturally

2 = correct response, but with slight, unnatural hesitation

1 = correct response, but given in a very slow, uncertain manner

0 = incorrect response or no response

. In answering the question, "¿Quiénes pagan por los comestibles?"

2 = correct use of the two articles required in responding

1 = correct use of one of the two articles required in responding

0 = incorrect use of both articles or no response.

. In describing the visual:

2 = correct verb and correct form

1 = correct verb but incorrect form

0 = incorrect verb and form or no response.

The scoring form for this test is shown in Appendix B.

Test Reliability. The composite reliability for the second semester test, as measured by the alpha coefficient, was:

$$\alpha = .401$$

This appeared to be a much more satisfactory figure than the ,740 obtained for the first test. It indicated, among other things, that the separate parts of the test were measuring the different aspects of speaking ability as intended. Intercorrelations of test parts and correlations of test parts with

the total speaking test and with the listening comprehension test are shown in table 4.

Table 4

CORRELATIONS BETWEEN SPEAKING TEST PARTS, TOTAL SPEAKING TEST, AND LISTENING COMPREHENSION TEST --SECOND SEMESTER OF 1960-61

Test Part	Structure	Fluency	Total Test	Listening Comprehension Test
Phonetic Accuracy	.180	.253	.236	.358
Structure		.419	.562	.088
Fluency			.951	.680
Total test				.713

Table 4, compared to table 1, shows a definite drop in intercorrelations of test parts. Correlations of test parts with the total test also went down, except for the fluency section, and, although the correlation of each part with the listening comprehension test was lower than before, the speaking test total correlated almost exactly the same (.713 vs. .714) with the listening comprehension test. Therefore, though this speaking test as a whole seemed to be measuring about the same skill as the first, each part was now doing its specific job more accurately.

The high correlation of the fluency section with the total test was disturbing, however. The fact that fluency did not correlate highly with the other test parts suggested that its high correlation with the total score was an artifact of the heavy weight given fluency in scoring. A pupil could get a maximum of 10 points on the phonetic accuracy section, two points on structure, and 16 points on fluency. The part-whole correlations support this explanation. For phonetic accuracy, the part-whole correlation with the total test was -.013; for structure, it was .396; and for fluency, it was .454. The drop from .951

to .454 on fluency indicates its heavy influence on the total score.

Rater reliabilities are given in table 5.

Table 5

RATER RELIABILITY ON THE
1960-61 SECOND SEMESTER SPEAKING TEST

Test Part	Reliability Coefficient r_{kk}	$\bar{r}_{11}$	r_{ab}
Phonetic Accuracy	.682	.517	.532
Structure	.893	.807	.807
Fluency	.989	.979	.980
Total Test	.984	.969	.971

Again, the rater reliabilities seemed highly satisfactory. Compared to the first semester test (shown in table 2), the reliabilities for the phonetic accuracy and structure sections went down a bit, while those for the fluency section and the total test went up.

Means and variances for the two raters on each test part are given in table 6. On this test, the differences between means were at chance level, which should be one result of the more specific definition of scale positions in scoring. Variances differed more than on the previous test, however, and this is reflected by the differences between $\bar{r}_{11}$ and r_{ab} in table 5. Even on phonetic accuracy, however, where the ratio of variances between raters was about three to two, $\bar{r}_{11}$ and r_{ab} differed by only .015; in most situations, these two rater reliability measures will apparently give about the same result if only two raters are used.

Table 6

MEANS AND VARIANCES FOR EACH RATER
ON THE 1960-61 SECOND SEMESTER SPEAKING TEST

Test Part	Rater	Mean	Variance
Phonetic Accuracy	A	9.020	1.059
	B	9.570	.654
Structure	A	1.210	.796
	B	1.005	.774
Fluency	A	3.880	12.426
	B	3.575	11.533
Total Test	A	14.110	20.160
	B	14.150	17.707

Table 6 raises a point about the usefulness of the phonetic accuracy section of the test. Since the maximum possible score was ten, the evaluators scored the average pupil about 93 per cent accurate on this part of the test, and the small variance shows that most pupils did this well. It was stated previously that the fluency section was unduly influencing the total score, yet table 6 indicates that it contributed only about a third as much as phonetic accuracy to the total. This apparent paradox is explainable through the high variance on fluency and very low variance on phonetic accuracy. Since each pupil was scoring about the same on phonetic accuracy, adding scores from this section to the total amounted to linearly transforming the total score. This would not affect the correlation of the total score with any other variable, and it would make the correlation of phonetic accuracy with the other test parts negligible. The part-whole correlation of phonetic accuracy with the total score was, in fact, -.013, which is not significantly different from zero.

Subsequent Tests

The statistical analysis of the second semester test showed that two further improvements were needed. The phonetic accuracy section needed revising

so that it would more accurately discriminate between pupils on their ability to pronounce Spanish sounds. This was accomplished in subsequent tests by eliminating the vowel "o," which is pronounced in Spanish the same way it is in English, by adding three or four of the difficult Spanish consonants, and by attempting to define the scale positions so that differences between excellent, fair, and poor pronunciation could be more accurately determined.

The structure section also needed a change. It worked fine per se, but its contribution to the total score was too small. This was overcome by using two sentences, that is, by doubling the size of the section, and by evaluating verb and adjective endings as well as syntax. (Vocabulary was established the same way as before, except for the precise verb and adjective forms to be used.) Finally, the weight of the fluency section was reduced by changing from a three- to a two-point scale, and this, of course, increased the relative weight of the structure section.

Sixth grade speaking tests were also developed. They employed the same general format and scoring procedure as the fifth grade tests, but they were built around sixth grade course content. Both fifth and sixth grade tests have been used in the 1961-62 and 1962-63 school years. At present, only the 1961-62 results have been analyzed. The tests have seemed to work well in every respect. Each part contributes about the same amount to the total score, intercorrelations among parts are low, and rater reliabilities, with three raters used, have averaged about .930. More important perhaps, different parts of the tests have revealed significant differences between methods of teaching elementary school Spanish.

Summary and Conclusions

Development of speaking skills is an important part of foreign language instruction, and adequate evaluation of a foreign language program depends in part on the measurement of speaking skills. The Denver-Stanford project is

involved with teaching Spanish to fifth and sixth grade pupils in the Denver Public Schools, and one of its concerns has been the evaluation of these pupils' abilities to speak Spanish. Since no tests of speaking ability at the elementary school level were available when the program started in 1960, project personnel began the development of speaking tests.

Careful review of FLES recommendations and of relevant literature led to the conclusion that the speaking skill could be broken down into three distinct aspects: the ability to pronounce Spanish sounds properly; the ability to structure Spanish sentences correctly; and the ability to communicate in Spanish with ease and naturalness. To measure these separate aspects -- phonetic accuracy, structure, and fluency -- speaking tests were constructed.

The tests were administered to random samples of pupils by project personnel. Each pupil's performance was recorded on magnetic tape, and each was in turn evaluated independently by at least two persons.

Statistical evaluation in the development process was limited entirely to internal validity and reliability. External validity was necessarily of the construct type since no outside criterion, against which to compare obtained results, was available. Both composite and rater reliabilities were computed. Since each part of the test measured a separate aspect of the speaking skill, internal validity would vary inversely with composite reliability. Consequently a low alpha coefficient, the measure of composite reliability, was sought. Rater reliability, on the other hand, reflected the extent to which similar scores were assigned each pupil by the evaluators, and, therefore, a high figure was sought. The three measures of rater reliability used were Ebel's mean for k ratings, r_{kk}, his mean for one rater, $\bar{r}_{11}$, and the Pearson product moment correlation coefficient, r_{ab}.

The development process revealed several points to be considered in constructing a foreign language speaking test. If the test parts are really

to reflect different aspects of the speaking skill, they must be evaluated separately, and the evaluator must be careful not to be influenced by performance on one section when scoring another. A five-point rating scale for specific test items, such as pronunciation of a vowel, demanded finer discrimination than the evaluators felt they could validly make, and a two- or three-point scale was found more desirable. Also, the evaluators felt they could make better judgments if each scale position was defined by a specific behavioral element. Finally, each test part should produce about the same mean score and about the same variance to weigh equally in the total test score.

The tests have been used in subsequent years and have been found satisfactory, both in terms of test criteria and in terms of differentiating between methods of teaching Spanish.

Appendix A

Speaking Test Form

Student _____

Research Group Assignment _____

I. Phonetic Accuracy

A	5	4	3	2	1	
E	5	4	3	2	1	
I	5	4	3	2	1	
O	5	4	3	2	1	
U	5	4	3	2	1	
n	5	4	3	2	1	
ll	5	4	3	2	1	Section 1 score _____

II. Structure

Sound (correct pronunciation)	5	4	3	2	1	
Order (syntax)	5	4	3	2	1	
Form (structure)	5	4	3	2	1	
Choice (vocabulary)	5	4	3	2	1	
Spontaneous response	5	4	3	2	1	
Appropriate answer	5	4	3	2	1	Section 2 score _____

III. Fluency

Expression of ideas	5	4	3	2	1	
Naturalness of utterances	5	4	3	2	1	
Vocabulary usage	5	4	3	2	1	
Sentence structure	5	4	3	2	1	Section 3 score _____

Total score _____

Appendix B

SPEAKING TEST SCORING FORM

REVISED

Student _____ School _____

Research Group Assignment _____ Student No. _____

Section A. Phonetic Accuracy

E	2	1	0
U	2	1	0
A	2	1	0
O	2	1	0
I	2	1	0

Section A. Score _____

Section B. Structure

Order (syntax)	2	1	0

Section B. Score _____

Section C. Fluency

1.	Appropriate answer	3	2	1	0
2.	Spontaneous response	3	2	1	0
3.	Accuracy in article agreement		2	1	0
4.	Accuracy in verb usage	(a)	2	1	0
		(b)	2	1	0
		(c)	2	1	0
		(d)	2	1	0

Section C. Score _____

TOTAL SCORE _____

Comments:

Appendix C

RELATIONSHIP OF $\bar{r}_{11}$ and r_{ab}

WHERE TWO RATERS ARE USED AND VARIANCE IS EQUAL

Consider the Ebel mean reliability coefficient of ratings for one rater, $\bar{r}_{11} = \frac{V_p - V_e}{V_p + V_e}$ and let ΣA = the sum of scores assigned by one rater and ΣB = the sum of scores assigned by the other rater. By definition,

$$V_p = \frac{\Sigma d_p^2}{df_p} = \frac{\frac{\Sigma(A+B)^2}{k} - \frac{(\Sigma A + \Sigma B)^2}{kN}}{(N-1)}$$

$$V_e = \frac{\Sigma d_e^2}{df_e} = \frac{\Sigma A^2 + \Sigma B^2 - \frac{\Sigma(A+B)^2}{k} - \frac{(\Sigma A)^2 + (\Sigma B)^2}{N} + \frac{(\Sigma A + \Sigma B)^2}{kN}}{(k-1)(N-1)}$$

where k = number of raters
N = number being rated

With k = 2,

$$\bar{r}_{11} = \frac{\frac{\Sigma(A+B)^2 - \frac{(\Sigma A + \Sigma B)^2}{N} - \Sigma A^2 - \Sigma B^2 + \frac{(\Sigma A)^2 + (\Sigma B)^2}{N}}{(N-1)}}{\frac{\Sigma A^2 + \Sigma B^2 - \frac{(\Sigma A)^2 + (\Sigma B)^2}{N}}{(N-1)}}$$

$$= \frac{\Sigma A^2 + \Sigma B^2 + 2\Sigma AB - \frac{(\Sigma A)^2 + (\Sigma B)^2 + 2\Sigma A \Sigma B}{N} - \Sigma A^2 - \Sigma B^2 + \frac{(\Sigma A)^2 + (\Sigma B)^2}{N}}{\Sigma A^2 + \Sigma B^2 - \frac{(\Sigma A)^2 + (\Sigma B)^2}{N}}$$

$$= \frac{2\Sigma AB - \frac{2\Sigma A \Sigma B}{N}}{[\Sigma A^2 - \frac{(\Sigma A)^2}{N}] + [\Sigma B^2 - \frac{(\Sigma B)^2}{N}]}$$

If the scores assigned by each rater have equal variance, then the two denominator terms are equal, and the equation reduces to:

$$\bar{r}_{11} = \frac{N\Sigma AB - \Sigma A\Sigma B}{[N\Sigma A^2 - (\Sigma A)^2]}$$

The Pearson product moment correlation coefficient between scores assigned by two raters is:

$$r_{ab} = \frac{\Sigma ab}{N\sigma_a\sigma_b}$$

By substitution, this reduces to the familiar computational formula:

$$r_{ab} = \frac{N\Sigma AB - \Sigma A\Sigma B}{\sqrt{N\Sigma A^2 - (\Sigma A)^2}\ \sqrt{N\Sigma B^2 - (\Sigma B)^2}}$$

If the scores assigned by each rater have equal variance, then the two denominator terms in this equation are also equal, and the product moment correlation reduces to:

$$r_{ab} = \frac{N\Sigma AB - \Sigma A\Sigma B}{[N\Sigma A^2 - (\Sigma A)^2]}$$

This was exactly the same result obtained when the Ebel coefficient was reduced under these conditions. Therefore, with two raters and equal variance of scores assigned by the raters, $\bar{r}_{11} = r_{ab}$.

References

Andrade, Manuel, John L. Hayman, Jr., and James T. Johnson, Jr. "Measurement of Listening Comprehension via Television in Elementary School Spanish Instruction." Denver-Stanford Project on the Context of Instructional Television. Report number 3. Denver, Colorado: Title VII Office, October, 1961. (Mimeo.)

Brooks, Nelson. Language and Language Learning. New York: Harcourt, Brace, and Company, 1960.

English, Horace B., and Ava Champney English. A Comprehensive Dictionary of Psychological and Psychoanalytical Terms. New York: Longmans, Green, and Company, 1958.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill Book Company, Inc., 1954.

Huebner, Theodore. How to Teach Foreign Languages Effectively. New York: New York University Press, 1959.

Keesee, Elizabeth. Modern Foreign Languages in the Elementary School. Washington, D. C.: U. S. Department of Health, Education, and Welfare, 1960.

MacRae, Margit W. Teaching Spanish in the Grades. Boston: Houghton Mifflin Company, 1957.

McNemar, Quinn. Psychological Statistics. New York: John Wiley and Sons, Inc., 1955.

The State Advisory Committee on Foreign Language Instruction. Foreign Languages Grades 7 - 12. Curriculum Bulletin Series No. V. Hartford, Connecticut: State Department of Education, September, 1958.